L7: Entry 4 of 4 File: USPT Dec 28, 2004

DOCUMENT-IDENTIFIER: US 6836773 B2
TITLE: Enterprise web mining system and method

Drawing Description Text (20):
FIG. 18 illustrates an example of inductive models generated using clustering and association algorithms.

Detailed Description Text (22):
Most data mining problems are addressed according to one of three paradigms: supervised learning, association analysis, and clustering. These paradigms have been applied to numerous problems in corporate and database mining such risk assessment, attrition and retention modeling, campaign marketing, fraud detection, customer profiling, profitability and cross-selling. These application problems are usually viewed from an account- or user-centric point of view. All the relevant information for each user is merged and consolidated in one record. An input dataset then looks like a large, mostly populated two-dimensional table where the columns correspond to attributes (independent variables). In the supervised learning approach, one particular column provides the `target` that is used as the dependent variable for the Data Mining model. Association modeling attempts to find associations: common patterns and trends in a less structured way (i.e. independent of a particular target field). These associations are supported by statistical correlations between different attributes of the dataset and are extracted by imposing independence, support, and confidence thresholds. Association analysis is applied to transaction or market basket data typically. In this case the datasets consists of transaction data listing a basket or group of items corresponding to an individual sale. The dataset is again a two-dimensional table but in this case potentially very sparse. Clustering is used for data-reduction and for class discovery. It is a method to find general correlation structures that group records into similarity groups. Clustering can be applied to both account or transaction-based datasets. Most data mining tool-sets support algorithms that provide instances of these paradigms but it is not common to encounter the three paradigms in a single problem.

Detailed Description Text (26):
Transform a web site visitor's behavior into data about his preferences.

Detailed Description Text (27):
Integrate web transactions and browsing behavior data with customer information and demographics

Detailed Description Text (29):
Help discover interesting and relevant patterns, clusters, and relationships in the transaction and user customer data.

Detailed Description Text (33):
Recommendations are personalized for each visitor to the Web site. This has distinct advantages over tailoring recommendations to broad, general market segments. Recommendations are based on a visitor's data and activity such as navigational behavior, ratings, purchases, as well as demographic data.

Detailed Description Text (36):
The personalization application works in conjunction with an existing Web application. The Web application asks the personalization application to record certain activities, and the data is saved by the personalization application into a schema. The Web application asks the personalization application to produce a list of products likely to be purchased by a Web site

visitor; a scored list of recommendations compiled from the visitor's current behavior and from data in another schema is passed to the Web application.

Detailed Description Text (42):
demographic data

Detailed Description Text (44):
Visitors to the Web site are of two types: registered visitors (customers) and unregistered visitors (visitors). For customers, the personalization application has both data from a current session and historical data collected over time for a given customer, as well as demographic data. For visitors, there is no historical data, so recommendations are based on current session behavior and demographic data, if available.

Detailed Description Text (49):
An exemplary block diagram of one embodiment of an enterprise web mining system 900, according to the present invention, is shown in FIG. 9. FIG. 9 is an example of physical and logical components that are combined to form the enterprise web mining system of the present invention. System 900 includes a plurality of data sources 902, a data preprocessing engine 903, a webhouse or web data warehouse 904, a web server 906, a data mining engine 908, a reporting engine 910, and web portal pages 912. Data sources 902 include corporate data 914, external data 916, Web transaction data 918, and Web server data 919. Corporate data 914 include the traditional proprietary corporate database or data warehouse that stores account- or user-based records. For example the name, age, amount of service or merchandise bought, length of time since initial creation, etc. External data 916 includes complementary data such as external demographics and other data acquired from external sources. Web transaction data 918 includes data relating to transactions, such as purchases, information requests, etc., which have been completed over the Web. Web data 919 includes Web traffic data from TCP/IP packet sniffing (live data collection), data obtained by direct access to the Web server's API, and Web server log files.

Detailed Description Text (52):
Data mining engine 908 may be based on any standard data mining technology, such as the ORACLE DARWIN 4.0.RTM. data mining engine. Data mining engine 908 generates data mining models using several machine learning technologies. Each machine learning technology is embodied in one or more modules that provide the model building functionality appropriate to each mode. Preferably, the supported machine learning technologies include: Naive Bayes modeling, Association rules, and decision tree models for the creation of inductive models. Naive Bayes models provide the capability of fast incremental learning. Decision trees of the classification and regression tree (CART) type provide transparent and powerful on-line rules and may be batch trained. In addition, a self organizing map clustering module provides the capability to address segmentation and profiling. The supported web mining methodologies provide the capability to perform a wide range of end-use functions. For example, the present invention may support the on-line customer lifecycle, which includes elements such as customer acquisition, customer growth, customer retention and lifetime profitability. Additional examples include click through optimization or web site organization.

Detailed Description Text (57):
Referring to FIG. 10, which is an exemplary data flow diagram of the methodological and technical framework of the enterprise web mining system 1000, implemented in the system shown in FIG. 9, system 1000 includes a plurality of data sources, such as corporate customer data 1002, which is typically provided by corporate database 914, complementary or external customer data 1004, which is typically provided by external databases 916, web server data 1006, which is typically provided by web database 919, and web transaction and visitor data 1008, which is typically provided by web transaction database 918. System 1000 includes a plurality of data processing blocks, such as feature selection and mapping blocks 1010 and 1012 and web data preprocessing block 1013, which are typically implemented in data preprocessing engine 903. System 1000 includes a plurality of data tables, such as account based table 1014, transaction based table 1016, and transaction summary table 1018, which are typically stored in webhouse 904. System 1000 includes a plurality of untrained data mining models, such as supervised learning model 1022, clustering model 1024, association model 1026, and statistical analysis model 1028, which are typically processed (trained) by data mining engine 908. System 1000 includes a plurality of trained data mining models, such as statistical summaries 1030, association rules 1032, clusters/segments 1034, and scoring models and rules 1036, as well as

reports, visualizations, scores and deployed models that are included in block 1040. The trained data mining models are typically processed by data mining engine 908, which generates the deployed models in block 1040. The deployed models are used by real time recommendation engine 924 to generate dynamic web pages, predictions, and recommendations 1042. The reports in block 1040 are typically generated by reporting engine 910. Other online processing is performed by online analytical processing (OLAP) engine 1038.

Detailed Description Text (60):
Step 1106 of process 1100 involves generating and deploying the models that are used to perform online recommendation and prediction. The processing of step 1106 is typically performed by data mining engine 908. Step 1106 includes a plurality of steps. Step 1106 begins with model setup step 1106-1, in which the algorithms that are to be used to generate the models are selected and setup. Once the algorithms and corresponding data structures are selected and setup, they may be viewed as untrained models, such as models 1022, 1024, 1026, and 1028. In step 1106-2, the representations that make up the trained models, such as information defining the logic, conditions, and decisions of the models, are generated using training data. These trained models may include statistical summaries 1030, association rules 1032, clusters/segments 1034, and scoring models and rules 1036. In step 1106-3, the representations of the generated models, such as blocks 1030, 1032, 1034, and 1036 of system 1000, are evaluated and refined to improve the quality of the model. In step 1106-4, the evaluated models are encoded in an appropriate format and deployed for use, such as in block 1040.

Detailed Description Text (64):
Data collection, step 1102 of process 1100, includes the acquisition 1102-1, selection 1102-2, pre-data mining processing of data 1102-3, and building of data tables 1102-4 that are to be used in the web mining process implemented in system 1000. Among the data sources that are utilized are corporate customer data 1002, complementary or external data 1004, Web server data 1006, and Web transaction and visitor data 1008. Corporate customer data 1002 includes the traditional corporate database or data warehouse that stores account- or user-based records. For example the name, age, amount of service or merchandise bought, length of time since initial creation, etc. Complementary data 1004 includes complementary data such as external demographics and other data acquired from external sources.

Detailed Description Text (107):
The types of models generated and used by the present invention may be categorized into several general classes. Among these classes are inductive models, supervised learning models, models using association and temporal pattern analysis, and models using clustering analysis.

Detailed Description Text (110):
The supervised learning algorithms used by the present invention include decision trees of the classification and regression tree (CART) type and Naive Bayes. CART is a very powerful non-parametric classification and regression method that produces accurate and easily interpretable models. It is a good representative of the wide class of decision-tree rule-based methods. A nice feature of decision-trees is the fact that the model is transparent, and can be represented as a set of rules in plain English, PL/SQL, Java or store procedures. This makes them ideal models for enterprise-wide business applications, query based analytical tools and e-commerce in general.

Detailed Description Text (115):
Clustering analysis is generally done in the context of class discovery, the finding of unknown groups or classes that define a taxonomy for the records at hand, or for data reduction by finding a small number of suitable representatives (centroids). In the present invention, clustering analysis algorithms include k-means and self-organizing maps (SOM) to provide the basic clustering. In addition to the algorithms, a method for cluster validation and interpretation (visualization) facilitates the use and evaluation of the results. The most important application to clustering is in the context of account-based tables, although transaction-based tables can also be clustered. Clustering can also be used to expose well-supported structure in the dataset and then to correlate this with a target class of interest. This amounts to a combined class discovery and interpretation methodology.

Detailed Description Text (171):
An example of an inductive model that uses clustering and associations is shown in FIG. 18. As an example of clustering, user and account data from table 1502 of FIG. 15, such as phone usage

data 1802, user age data 1804, and calling card usage data 1806 is analyzed to located clusters of data that may be modeled. As an example of association, session data from table 1508 of FIG. 15, such as whether the user clicked on the modems link 1808 and whether the user visited the products page, and keyword data from table 1514 of FIG. 15, such as searching on the keyword "computer" 1812, is analyzed to determine associations among data that may be modeled.

Detailed Description Text (185):
Decision trees and association rules return recommendations based on abstractions (models) of shopping cart history or corporate records that are built in advance. K-nearest neighbors score the current shopping cart against the table of aggregate transactions for each customer. Confidence measure for each possible recommended product can be constructed for all three methods. These confidence measures should be complemented with weights derived from business rules. For example, although product A is a product more likely to be bought than B, the profit from product B is higher, making it a more desirable product to be sold from the merchant's point of view. The key measure is the expected profit from a recommendation: (probability (confidence) of a recommendation being bought).times.profit. Here is a clear example of why an application-oriented layer is necessary. In the third case above where all the different tables are used, a two-stage process is probably desirable. First the customer profile is recovered by assigning him to a demographic and a browsing behavior cluster. Then the recommendation is computed taking in account only the transactions generated from customers belonging to the same profile. The rational here is that we should look for similar basket among people with similar demographics, for example.

Detailed Description Text (189):
Segmentation can be done using the profiling clusters or the un-clustered customer data. The first is quick and allows many different studies to be quickly performed. The un-clustered customer data case is slower but probably more precise. In the case of segmentation a measurement has to be selected. For example: purchases in dollar can be used to segment customers (or clusters) into bad, average, good customers.

Previous Doc Next Doc Go to Doc#

L7: Entry 3 of 4 File: PGPB Jun 27, 2002

DOCUMENT-IDENTIFIER: US 20020083067 A1
TITLE: Enterprise web mining system and method

Brief Description of Drawings Paragraph:
[0033] FIG. 18 illustrates an example of inductive models generated using clustering and association algorithms.

Detail Description Paragraph:
[0059] Most data mining problems are addressed according to one of three paradigms: supervised learning, association analysis, and clustering. These paradigms have been applied to numerous problems in corporate and database mining such risk assessment, attrition and retention modeling, campaign marketing, fraud detection, customer profiling, profitability and cross-selling. These application problems are usually viewed from an account- or user-centric point of view. All the relevant information for each user is merged and consolidated in one record. An input dataset then looks like a large, mostly populated two-dimensional table where the columns correspond to attributes (independent variables). In the supervised learning approach, one particular column provides the `target` that is used as the dependent variable for the Data Mining model. Association modeling attempts to find associations: common patterns and trends in a less structured way (i.e. independent of a particular target field). These associations are supported by statistical correlations between different attributes of the dataset and are extracted by imposing independence, support, and confidence thresholds. Association analysis is applied to transaction or market basket data typically. In this case the datasets consists of transaction data listing a basket or group of items corresponding to an individual sale. The dataset is again a two-dimensional table but in this case potentially very sparse. Clustering is used for data-reduction and for class discovery. It is a method to find general correlation structures that group records into similarity groups. Clustering can be applied to both account or transaction-based datasets. Most data mining tool-sets support algorithms that provide instances of these paradigms but it is not common to encounter the three paradigms in a single problem.

Detail Description Paragraph:
[0063] Transform a web site visitor's behavior into data about his preferences.

Detail Description Paragraph:
[0064] Integrate web transactions and browsing behavior data with customer information and demographics

Detail Description Paragraph:
[0066] Help discover interesting and relevant patterns, clusters, and relationships in the transaction and user customer data.

Detail Description Paragraph:
[0070] Recommendations are personalized for each visitor to the Web site. This has distinct advantages over tailoring recommendations to broad, general market segments. Recommendations are based on a visitor's data and activity such as navigational behavior, ratings, purchases, as well as demographic data.

Detail Description Paragraph:
[0073] The personalization application works in conjunction with an existing Web application. The Web application asks the personalization application to record certain activities, and the data is saved by the personalization application into a schema. The Web application asks the personalization application to produce a list of products likely to be purchased by a Web site visitor; a scored list of recommendations compiled from the visitor's current behavior and from

data in another schema is passed to the Web application.

Detail Description Paragraph:
[0079] demographic data

Detail Description Paragraph:
[0081] Visitors to the Web site are of two types: registered visitors (customers) and unregistered visitors (visitors). For customers, the personalization application has both data from a current session and historical data collected over time for a given customer, as well as demographic data. For visitors, there is no historical data, so recommendations are based on current session behavior and demographic data, if available.

Detail Description Paragraph:
[0086] An exemplary block diagram of one embodiment of an enterprise web mining system 900, according to the present invention, is shown in FIG. 9. FIG. 9 is an example of physical and logical components that are combined to form the enterprise web mining system of the present invention. System 900 includes a plurality of data sources 902, a data preprocessing engine 903, a webhouse or web data warehouse 904, a web server 906, a data mining engine 908, a reporting engine 910, and web portal pages 912. Data sources 902 include corporate data 914, external data 916, Web transaction data 918, and Web server data 919. Corporate data 914 include the traditional proprietary corporate database or data warehouse that stores account- or user-based records. For example the name, age, amount of service or merchandise bought, length of time since initial creation, etc. External data 916 includes complementary data such as external demographics and other data acquired from external sources. Web transaction data 918 includes data relating to transactions, such as purchases, information requests, etc., which have been completed over the Web. Web data 919 includes Web traffic data from TCP/IP packet sniffing (live data collection), data obtained by direct access to the Web server's API, and Web server log files.

Detail Description Paragraph:
[0089] Data mining engine 908 may be based on any standard data mining technology, such as the ORACLE DARWIN 4.0.RTM. data mining engine. Data mining engine 908 generates data mining models using several machine learning technologies. Each machine learning technology is embodied in one or more modules that provide the model building functionality appropriate to each mode. Preferably, the supported machine learning technologies include: Naive Bayes modeling, Association rules, and decision tree models for the creation of inductive models. Naive Bayes models provide the capability of fast incremental learning. Decision trees of the classification and regression tree (CART) type provide transparent and powerful on-line rules and may be batch trained. In addition, a self organizing map clustering module provides the capability to address segmentation and profiling. The supported web mining methodologies provide the capability to perform a wide range of end-use functions. For example, the present invention may support the on-line customer lifecycle, which includes elements such as customer acquisition, customer growth, customer retention and lifetime profitability. Additional examples include click through optimization or web site organization.

Detail Description Paragraph:
[0094] Referring to FIG. 10, which is an exemplary data flow diagram of the methodological and technical framework of the enterprise web mining system 1000, implemented in the system shown in FIG. 9, system 1000 includes a plurality of data sources, such as corporate customer data 1002, which is typically provided by corporate database 914, complementary or external customer data 1004, which is typically provided by external databases 916, web server data 1006, which is typically provided by web database 919, and web transaction and visitor data 1008, which is typically provided by web transaction database 918. System 1000 includes a plurality of data processing blocks, such as feature selection and mapping blocks 1010 and 1012 and web data preprocessing block 1013, which are typically implemented in data preprocessing engine 903. System 1000 includes a plurality of data tables, such as account based table 1014, transaction based table 1016, and transaction summary table 1018, which are typically stored in webhouse 904. System 1000 includes a plurality of untrained data mining models, such as supervised learning model 1022, clustering model 1024, association model 1026, and statistical analysis model 1028, which are typically processed (trained) by data mining engine 908. System 1000 includes a plurality of trained data mining models, such as statistical summaries 1030, association rules 1032, clusters/segments 1034, and scoring models and rules 1036, as well as reports, visualizations, scores and deployed models that are included in block 1040. The

trained data mining models are typically processed by data mining engine 908, which generates the deployed models in block 1040. The deployed models are used by real time recommendation engine 924 to generate dynamic web pages, predictions, and recommendations 1042. The reports in block 1040 are typically generated by reporting engine 910. Other online processing is performed by online analytical processing (OLAP) engine 1038.

Detail Description Paragraph:
[0097] Step 1106 of process 1100 involves generating and deploying the models that are used to perform online recommendation and prediction. The processing of step 1106 is typically performed by data mining engine 908. Step 1106 includes a plurality of steps. Step 1106 begins with model setup step 1106-1, in which the algorithms that are to be used to generate the models are selected and setup. Once the algorithms and corresponding data structures are selected and setup, they may be viewed as untrained models, such as models 1022, 1024, 1026, and 1028. In step 1106-2, the representations that make up the trained models, such as information defining the logic, conditions, and decisions of the models, are generated using training data. These trained models may include statistical summaries 1030, association rules 1032, clusters/segments 1034, and scoring models and rules 1036. In step 1106-3, the representations of the generated models, such as blocks 1030, 1032, 1034, and 1036 of system 1000, are evaluated and refined to improve the quality of the model. In step 1106-4, the evaluated models are encoded in an appropriate format and deployed for use, such as in block 1040.

Detail Description Paragraph:
[0101] Data collection, step 1102 of process 1100, includes the acquisition 1102-1, selection 1102-2, pre-data mining processing of data 1102-3, and building of data tables 1102-4 that are to be used in the web mining process implemented in system 1000. Among the data sources that are utilized are corporate customer data 1002, complementary or external data 1004, Web server data 1006, and Web transaction and visitor data 1008. Corporate customer data 1002 includes the traditional corporate database or data warehouse that stores account- or user-based records. For example the name, age, amount of service or merchandise bought, length of time since initial creation, etc. Complementary data 1004 includes complementary data such as external demographics and other data acquired from external sources.

Detail Description Paragraph:
[0144] The types of models generated and used by the present invention may be categorized into several general classes. Among these classes are inductive models, supervised learning models, models using association and temporal pattern analysis, and models using clustering analysis.

Detail Description Paragraph:
[0147] The supervised learning algorithms used by the present invention include decision trees of the classification and regression tree (CART) type and Naive Bayes. CART is a very powerful non-parametric classification and regression method that produces accurate and easily interpretable models. It is a good representative of the wide class of decision-tree rule-based methods. A nice feature of decision-trees is the fact that the model is transparent, and can be represented as a set of rules in plain English, PL/SQL, Java or store procedures. This makes them ideal models for enterprise-wide business applications, query based analytical tools and e-commerce in general.

Detail Description Paragraph:
[0152] Clustering analysis is generally done in the context of class discovery, the finding of unknown groups or classes that define a taxonomy for the records at hand, or for data reduction by finding a small number of suitable representatives (centroids). In the present invention, clustering analysis algorithms include k-means and self-organizing maps (SOM) to provide the basic clustering. In addition to the algorithms, a method for cluster validation and interpretation (visualization) facilitates the use and evaluation of the results. The most important application to clustering is in the context of account-based tables, although transaction-based tables can also be clustered. Clustering can also be used to expose well-supported structure in the dataset and then to correlate this with a target class of interest. This amounts to a combined class discovery and interpretation methodology.

Detail Description Paragraph:
[0211] An example of an inductive model that uses clustering and associations is shown in FIG. 18. As an example of clustering, user and account data from table 1502 of FIG. 15, such as

phone usage data 1802, user age data 1804, and calling card usage data 1806 is analyzed to located clusters of data that may be modeled. As an example of association, session data from table 1508 of FIG. 15, such as whether the user clicked on the modems link 1808 and whether the user visited the products page, and keyword data from table 1514 of FIG. 15, such as searching on the keyword "computer" 1812, is analyzed to determine associations among data that may be modeled.

Detail Description Paragraph:
[0235] Decision trees and association rules return recommendations based on abstractions (models) of shopping cart history or corporate records that are built in advance. K-nearest neighbors score the current shopping cart against the table of aggregate transactions for each customer. Confidence measure for each possible recommended product can be constructed for all three methods. These confidence measures should be complemented with weights derived from business rules. For example, although product A is a product more likely to be bought than B, the profit from product B is higher, making it a more desirable product to be sold from the merchant's point of view. The key measure is the expected profit from a recommendation: (probability (confidence) of a recommendation being bought).times.profit. Here is a clear example of why an application-oriented layer is necessary. In the third case above where all the different tables are used, a two-stage process is probably desirable. First the customer profile is recovered by assigning him to a demographic and a browsing behavior cluster. Then the recommendation is computed taking in account only the transactions generated from customers belonging to the same profile. The rational here is that we should look for similar basket among people with similar demographics, for example.

Detail Description Paragraph:
[0239] Segmentation can be done using the profiling clusters or the un-clustered customer data. The first is quick and allows many different studies to be quickly performed. The un-clustered customer data case is slower but probably more precise. In the case of segmentation a measurement has to be selected. For example: purchases in dollar can be used to segment customers (or clusters) into bad, average, good customers.

Previous Doc Next Doc Go to Doc#